Modern Linguistics 现代语言学, 2023, 11(1), 8-16
Published Online January 2023 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2023.111002

Hans汉斯

基于语料库的人文社科学术论文词块功能研究

赵子祺，王雪霏

北京工业大学文法学部，北京

收稿日期：2022年11月24日；录用日期：2022年12月29日；发布日期：2023年1月5日

摘　要

本文基于MICUSP语料库，聚焦人文社会科学领域三个不同学科(语言学、经济学、社会学)的学术论文三词词块，从语言背景和学科视角，旨在对比研究英语本族语和非本族语者三词词块功能使用差异，以及三个学科间的三词词块功能分布是否存在相似性。结果表明，非本族语者与本族语者相比，产出的各功能类别词块数目相似，但使用的词块功能分布频率呈显著性差异；此外，同一领域学科的学术论文词块功能侧重点不尽相同。语言学倾向于使用“结果”等“文本导向型”词块，而经济学和社会学擅长使用“过程”和“描述”等“研究导向型”词块。

关键词

词块，语料库，词块功能，人文社会科学

A Corpus-Based Study on the Functions of Chunks in Academic Papers in Humanities and Social Sciences

Ziqi Zhao, Xuefei Wang

Faculty of Humanities and Social Sciences, Beijing University of Technology, Beijing

Received: Nov. 24th, 2022; accepted: Dec. 29th, 2022; published: Jan. 5th, 2023

Abstract

Based on MICUSP (Michigan Corpus of Upper-level Student Papers), this paper focuses on the three-word chunks extracted from academic papers across three different disciplines (linguistics,

文章引用: 赵子祺, 王雪霏. 基于语料库的人文社科学术论文词块功能研究[J]. 现代语言学, 2023, 11(1): 8-16.
DOI: 10.12677/ml.2023.111002

economics, sociology) of the humanities and social sciences. This paper aims to compare the functional use of three-word chunks produced by English native speakers and non-native speakers, and explore whether different disciplines in the same field use chunks in a functionally similar way. The results show that under each functional category, non-native speakers can produce similar number of chunks with native speakers, but there is a significant difference in the functional distribution frequency of the chunks they use. Besides, disciplines in the same field put different emphasis on the functions of chunks. Linguistics prefers to use "text-oriented" chunks, especially "resultative signals", while economics and sociology are apt to use "research-oriented" chunks like "procedure" and "description" chunks.

Keywords

Chunk, Corpus, Functions of Chunks, Humanities and Social Sciences

1. 引言

词块在语言交际活动中具有重要意义。自 19 世纪 80 年代起，借助语料库这一包含大量真实语言材料的大规模电子文本库，词块的相关研究得到了极大的促进发展。这类研究揭示了成人语言中普遍存在着一些“预制复现”的词汇模式或词块[1] [2]，且可以作为现成的记忆构件来使用[3]。这些多词词块无需经过语法规则处理就可以整体提取[2]，可以帮助提高理解速度和输出流畅性[4]。

自从语料库语言学出现以来，越来越多的研究者开始对比研究不同群体的词块使用差异。此类对比研究通常从语言背景、体裁和学科三个视角进行。以往研究发现，非本族语者的词块使用数量和多样性都亚于本族语者[5] [6]，两类群体对词块结构的使用偏好也不尽相同[7]；不同体裁依赖于不同类型的词块[8] [9]；硬学科和软学科领域对词块结构和功能的偏好存在明显差异[10]，且不同学科间存在核心词块[11]。

以往多数研究分析了作者水平、语言背景、学科领域因素对词块使用的影响，但很少探究同一领域的不同学科所使用的词块功能是否也存在差异，使用者的语言背景是否仍会影响词块功能使用偏好。而且以往研究多以四词词块为对象[10] [12] [13] [14]，对三词词块的研究则相对较少。但三词词块较四词和五词词块而言能提供更充分的数据，可以更好反映写作者的语言使用特点和倾向。

本文将借助 MICUSP 语料库，探讨人文社会科学领域三个不同学科(语言学、经济学、社会学)的学术论文中三词词块功能用法是否也具有高度的学科依赖性，以及本族语和非本族语者使用这些词块功能的方式是否存在差异。

2. 研究方法

2.1. 词块功能框架

词块研究通常基于某一分类框架进行。Cortes 提出了功能分类框架，包含四个类别：指示标记、文本组织、立场和学科特定词块[13]。Biber 等学者根据词块在语境中的功能，将词块分为立场、组篇和指示词块[9]。Hyland [10]修改了 Biber 的分类框架，使其更适于对学术研究中的词块进行分类。经比较评估，本研究采用 Hyland 的词块功能分类[10] (见表 1)。

Table 1. Hyland's framework of chunk functions
表 1. Hyland 词块功能框架

功能范畴	功能类型	例子
研究导向型	时间/位置	at the beginning of, at the same time, in the present study
	过程	the use of the, the role of the, the purpose of
	量化	the magnitude of, a wide range of, one of the most
	描述	the structure of the, the size of the, the surface of the
	话题	in the Hong Kong, the currency board system
文本导向型	过渡	on the other hand, in addition to the, in contrast to the
	结果	as a result of, it was found that, these results suggest that
	结构	in the next section, as shown in figure, as shown in table
	框架	in the case of, with respect to the, on the basis of, in the presence of
参与者导向型	立场	are likely to be, may be due to, it is possible that
	介入	it should be noted, as can be seen

在这一框架中，“研究导向型”词块描述研究相关的内容，如介绍实验过程结果；“文本导向型”词块用来组织语篇，如承接上文、指引读者阅读文章；“参与者导向型”词块关注文本的作者或读者，涉及到作者与读者的互动，如传达作者的判断。

2.2. 研究问题

本文将回答以下两个研究问题：

1) 就词块功能而言，经济学、语言学与社会学的本族语者与非本族语者学术论文中使用的三词词块是否存在差异？

2) 这三个学科在学术论文中的三词词块功能使用是否相似？

3) 若三个学科的三词词块功能不相似，这些词块在功能维度上的差异有多大？

2.3. 研究对象

研究数据由 5 个自建语料库组成，共计 149,417 个单词。语料收集自密歇根优秀学生论文语料库(MICUSP)。此语料库是由密歇根大学开发的书面英语语料库，由 16 个学科的本科四年级学生和一到三年级研究生所写的 830 篇 A 级论文组成。论文涵盖七个种类，包括报告、研究性论文、议论文、批判性论文等。学生的语言背景分为两类，一类为英语本族语者，另一类为非英语本族语者，包含以汉语、俄语、西班牙等语言为母语的学生。五个自建语料库包括本族语者(NS)语料库、非本族语者(NNS)语料库、经济学(ECO)语料库、语言学(LIN)语料库和社会学(SOC)语料库。前两个语料库分别包含本族语者和非本族语者撰写的论文，不对学科作区分。后三个语料库分别包含各科的学术论文，不对作者语言背景作区分。所选的学生水平包括一到三年级研究生。经济学、语言学及社会学在 MICUSP 中有相对充足的同类型论文，出于对语料可比性的考虑，本文选择这三个学科进行研究(见表 2)。

Table 2. Composition of the corpora
表 2. 语料库语篇类型分布情况

语料库	报告	研究论文	论文提案	语料库规模(形符数)
NS 语料库	11	10	5	78,286
NNS 语料库	10	8	4	71,131
ECO 语料库	3	9	2	44,738
LIN 语料库	6	6	2	54,281
SOC 语料库	12	3	5	50,362

2.4. 研究工具

选取文本首先经 TextEditor 3.0 进行文本清洁处理，再手动删除了除正文以外的标题、副标题、参考文献、表格和公式等部分。Antconc 3.5.9 语料库检索工具具有词表生成、主题词计算、搭配和词族提取等多种功能，本文借助其 n-gram 功能提取所有的三词词块。将词块筛选后，本文使用 UCREL 显著性检验系统对各类词块数目进行了对数似然检验，来找出不同组作者的词块使用倾向的统计显著性差异。

2.5. 数据收集

Biber 和 Barbieri 认为词块必须出现在语料库一定数量的文本中才能称为词块，如三到五个文本[15]。而 Chen 和 Baker 将词块的划分标准设定为至少出现在三个文本中[14]。对于频率的标准，大部分词块研究采用的频率阈值为每百万单词出现至少 10 次[12] 992 到每百万单词出现至少 40 次[16]不等。本研究采取较稳妥的词块提取标准，即每百万词至少出现 40 次，且至少出现在 5 个文本中。

按照此标准，NS 语料库中提取出 117 个三词词块，NNS 语料库 89 个，ECO、LIN、SOC 语料库的产出量分别为 40、57、52 个词块。卫乃兴将词块定义为语料库中高频出现的不同长度的有意义的连续词语片段，此类词语片段一般结构良好且意义相对完整[17]。根据此标准，作者手动删除了一些意义不完整的非结构化词块，以保证词块的研究价值，如“is that the”和“fact that the”。此外，如“it is a”、“there is the”、“it is the”等词块，单独来看没有完整的意义，很难按照 Hyland 的功能框架进行分类，因此这些词块也被删除。最后，NS 语料库中剩下 79 个词块，NNS 语料库中剩下 57 个。ECO、LIN、SOC 语料库中分别剩余 30、34、39 个词块。

每个语料库中的剩余词块根据 Hyland 的词块功能框架进行分类，这一过程由两人分别完成，采用简单百分比一致性法对评测者间信度进行检验，结果为 85.0%，证明评测者间误差较小，分类结果较为可靠。表 3 和表 4 列出了这些词块的功能分类和分布频率信息。

Table 3. Distribution of types and frequency of chunk in NS and NNS corpora
表 3. NS 及 NNS 语料库中词块的功能分类和分布频率

功能范畴	类别		频率	
	NS	NNS	NS	NNS
研究导向型	44	35	406	374*
文本导向型	27	18	298	225*
参与者导向	8	4	75	26*
总计	79	57*	779	625*

注：*$p < 0.05$。

Table 4. Functional distribution of chunk frequency in subcategories
表 4. 三个学科语料库的词块功能分布频率

功能范畴	功能类型	ECO	LIN	SOC
研究导向型	时间/位置	7	7	7
	过程	37	19*	31
	量化	50	64	9*
	描述	124	50*	119
	话题	7*	0	0
	总计	225*	140*	166*
文本导向型	过渡	21	25	60*
	结果	12	29*	13
	结构	23	30	0*
	框架	68	91	28*
	总计	124	175	101*
参与者导向型	立场	52*	22*	8*
	介入	0	8*	0
	总计	52*	30*	8*
总计		401*	345	275

注：*$p < 0.05$。

3. 数据分析和讨论

3.1. NS 及 NNS 语料库词块功能使用对比

由表 3 可以看出，虽然两库各功能类别的词块分布频率呈显著性差异，但各功能类别词块数目并无显著性差异。由此可以推断，非本族语学生与本族语学生能够产出的各功能类别词块数目相近，本研究中所涉及的非本族语学生能够较为熟练地运用各功能词块。但本族语学生在论文中会更频繁地使用词块，这一倾向同时反映在三种功能类别上。学习者水平越高越会依赖于使用词块[10]，然而，虽然本研究涉及的非本族语学生是水平相对较高的研究生，但他们对词块的依赖性仍低于本族语学生。因此，词块学习对高水平非本族语学生来说仍然是必要的。

两组学习者在“参与者导向型”词块的频率差异最大。这类词块表示作者的介入，如表达自己的立场。一些作者会避免使用这些词，以减少学术论文中的主观性。由表 3 可看出非本族语者使用“参与者导向型”词块的频率远低于本族语者，这种回避反映了语言背景因素对词块使用的影响。由于教育经历和文化偏好，一些非本族语作者不习惯于表达自己的立场[10]。由此可推断，尽管两库包含了同一领域相似学科的论文，但语言背景因素仍会影响非本族语作者对“参与者导向型”词块的使用。

3.2. 三个学科语料库间词块功能分布对比

根据 Hyland 的功能分类框架中的子类别，本文将三个学科语料库中的词块进行了更详细的分类，词

块功能分布频率如表 4 所示。为探究三个语料库间各功能的词块分布频率差异是否具有统计学意义，本文采用三路比较法(ECO/LIN、ECO/SOC、LIN/SOC)，使用对数似然统计量检验词块功能分布频率差异。如果一个学科某一分类与其他两个学科同一分类相比，词块分布频率都具有显著性差异($p < 0.05$)，那么这一分类组数据则标记为具有显著性差异。

3.2.1. 功能范畴对比

如图 1 所示，在 LIN 语料库中，“文本导向型”词块的频率多于“研究导向型”词块，而 ECO 和 SOC 语料库中情况相反。“文本导向型”词块利用语言强调衔接、指明情况和表明限制，从而证明观点。高比例的“文本导向型”词块能展示出作者一定的语言意识，表现出作者在关注研究的同时，还重视了语言组织[18]。相比经济学和社会学的学生，语言学专业的学生会更多地考虑到语言，因此会更倾向于使用“文本导向型”词块。相比而言，社会学专业的学生较少使用“文本导向型”词块，造成这种差异的因素可能是文本类型的数量不平衡。SOC 语料库中只有 3 篇研究论文，但有 12 篇报告。“文本导向型”词块有助于在长篇论文中连贯地传递论点，由于报告通常包含比研究论文更少的章节，因此不需要大量的“文本导向型”词块来组织语篇。

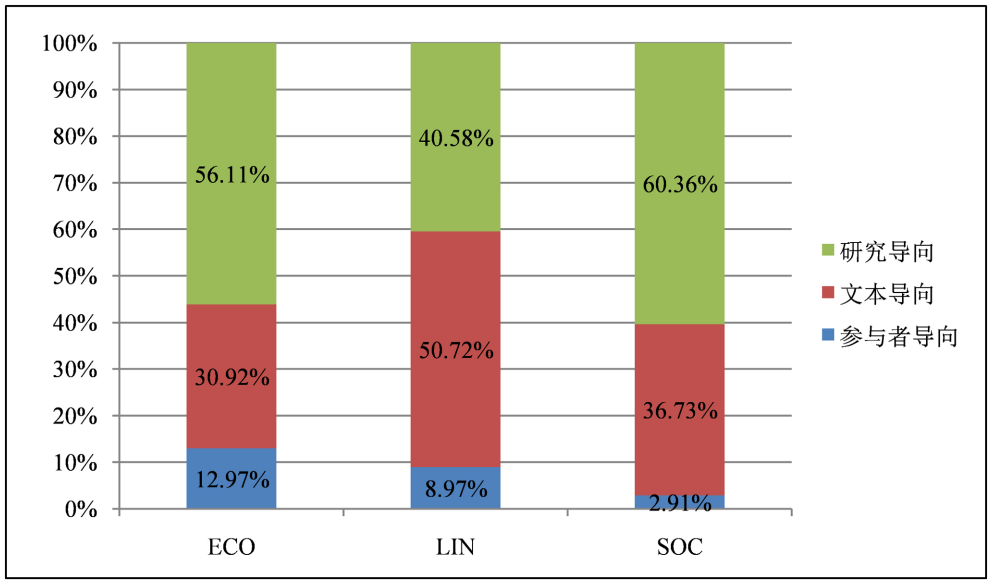

Figure 1. Functional distribution (frequency) of chunk in three corpora
图 1. 三个学科词块功能分布频率占比

在 Hyland [10]的研究中，硬学科(电子工程、生物学)作者倾向于在实证论证和实验结果的基础上使用“研究导向型”词块来表达自己的观点。本文三个学科虽同属于人文社科领域，但语言学偏向于人文学科，而经济学和社会学偏向于社会科学，社会科学是用科学的方法研究人类社会，是人文科学与自然科学之间的中介学科。因此，经济学和社会学比语言学更注重科学性，更多地使用“研究导向型”词块，介绍论文中的研究设计和过程。

从表 4 可以看出，“参与者导向型”词块在三个语料库中所占比例均为最低，与以往的大多数研究结果相符。经济学使用的“参与者导向型”词块比预期多，而社会学使用得比预期少，这一发现也可以用学科特征来解释。经济学以人是理性的这一假设为基础，研究这些理性的人如何相互作用以实现利润最大化，且人与人之间的互动是复杂的，因此经济学论文的结论会受多种因素的影响。“参与者导向型”

词块的使用表示作者不愿选择介入性的个人声音，为保证论文中的论述不会太主观绝对，经济学学生便使用“are likely to”等“参与者导向型”词块来引出论文中的猜测性论述。而社会学视社会上的人为一个整体，关注并研究这个整体的社会生活。从这个角度来看，社会学学生会倾向于通过描述事实来展开观点，因此会较少使用“参与者导向型”词块表达态度或提出假设。

3.2.2. 功能子类对比

语言学作为一门人文学科，并不像经济学和社会学那样重视厘清研究过程，因此 LIN 语料库中“过程”和“描述”的词块频率与 ECO 和 SOC 语料库在统计学上存在差异。经济学和社会学作者在论文中较多使用实证方法，因此他们倾向于使用“研究导向型”分类下的词块来展示他们开展研究的能力。

文本导向范畴下的“结果型”词块越多，作者的读者意识越高，因为这一词块表示作者对研究结果的解释，并突出了作者希望读者得出的推论[18]。某种程度上作者的读者意识与语言意识有关，因为语言组织得越好，越方便读者理解文章。因此，在 LIN 语料库中，“结果型”词块频率高也反映了语言学作者的语言学科能力。

SOC 语料库中没有“结构型”词块，这可能是因为：一是语料库规模太小，包含的文本数量有限，无法提取出“结构型”词块；第二个因素是文本的长度，较长的文本通常使用更多的“结构型”词块来引导读者来顺畅地阅读文本[16]，而 SOC 语料库中的文本平均长度(2518 个形符)比其他两个语料库的文本平均长度短(3196 个形符和 3877 个形符)，因此 SOC 语料库中“结构型”词块频率与其他两个库相比存在显著差异。“框架型”词块对读者的引导作用与“结构型”词块相似，因此“框架型”词块在 SOC 语料库中的频率也较低。

“介入型”词块是指以读者为中心的词块[10]。由于相较于其他两个学科，语言学作者的读者意识更高，因此 LIN 语料库中“介入型”词块的频率也相对更高。

3.2.3. 共用核心词块对比

三个语料库共同使用了 7 个词块(见表 5)，分别占 LIN、ECO、SOC 语料库的 23.33%、20.59%和 17.9%。以往有研究认为学术语篇中存在核心词块[11]，本研究在一定程度上证明了至少人文领域的学术论文中存在不同学科作者所共用的核心词块。这些核心词块分布在“研究导向型”和“文本导向型”类别中，用于介绍研究或组织语篇，不直接表述作者观点，相对来说具有一定客观性。因此，核心词块在学术论文中更有可能发挥文本导向或研究导向的作用，在文章中呈现客观性。

Table 5. Functional distribution of shared chunks
表 5. 共用词块功能分布

词块	功能范畴
as well as	文本导向型
based on the	文本导向型
be able to	研究导向型
in order to	研究导向型
is based on	文本导向型
part of the	研究导向型
the relationship between	研究导向型

4. 结论

本研究通过对比 NS 和 NNS 语料库，发现两者各功能类别的词块产出数量没有明显差异，但 NS 语料库中各功能类别的词块分布频率更高。本研究结果呼应了以往研究，即非本族语学生对词块依赖性远低于本族语学生。先前的一项研究表明，语言因素可能会对“参与者导向型”词块的使用产生影响，本研究进一步证明，这种影响在不同语言背景和相似学科的学术论文中仍然起作用。

ECO、LIN 和 SOC 语料库提取出 7 个共用词块，占每个语料库提取词块总数的近 20%，这验证了以往研究中关于核心词块的表述。同一领域的不同学科的学术论文会共用一些词块，但不同学科对词块功能偏好不同，这一偏好是由研究方法和研究对象等学科特点决定的。经济学和社会学是更注重科学方法的社会科学类学科，此类学术论文倾向于使用“过程”和“描述”等“研究导向型”词块来叙述研究相关方法和过程等内容。语言学作者具有更高的语言意识，因此会倾向于使用“结果”等“文本导向型”词块。经济学比社会学使用更多的“参与者导向型”词块，因为经济学研究中存在大量不确定性因素，这些因素带来的假设需要更多的“参与者导向型”词块，从而在文中加入作者观点，如 it is likely，来表明作者的猜测。由于语言意识与作者的读者意识有关，语言学学生也倾向于使用以读者为中心的“介入型”词块。

然而，由于 MICUSP 中文本数量有限，本研究建立的语料库规模较小，因此可能导致一些常用词块未达到提取频率，使得本文提取到的词块与真实使用情况存在偏差。另外，文本类型的不均衡也可能导致了某些语料库提取出的词块功能分布存在偏差。

最后，本研究对非本族语者的学术写作教学也有一定的指导意义。首先，教师应该向学生强调某些类别词块的使用，尤其是“参与者导向型”词块。非本族语作者可能片面地认为使用“参与者导向型”词块会增加文章的主观性，而降低学术论文的客观性。然而，在学术论文写作中，非本族语作者应该像本族语作者一样，善于利用“参与者导向型”词块来传递自己的观点，来与读者进行互动交流。此外，虽然同一领域的学科共用一些词块，但不同学科会偏好不同功能的词块。因此，教师在教授不同学科的学生时，应提高学生使用相应功能类别词块的意识，在教学中侧重介绍具有学科导向的词块。

参考文献

[1] Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford University Press, Oxford, 108.

[2] Lewis, M. (1993) The Lexical Approach: The State of ELT and a Way Forward. Language Teaching Publications, Hove.

[3] Pawley, A. and Syder, F.H. (1983) Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In: Richards, J.C. and Schmidt, R.W., Eds., *Language & Communication*, Longman, Harlow, 191-226.

[4] Siyanova-Chanturia, A., Conklin, K. and Schmitt, N. (2011) Adding More Fuel to the Fire: An Eye-Tracking Study of Idiom Processing by Native and Non-Native Speakers. *Second Language Research*, **29**, 72-89. https://doi.org/10.1177/0267658310382068

[5] Ädel, A. and Erman, B. (2012) Recurrent Word Combinations in Academic Writing by Native and Non-Native Speakers of English: A Lexical Bundles Approach. *English for Specific Purposes*, **31**, 81-92. https://doi.org/10.1016/j.esp.2011.08.004

[6] Howarth, P. (1998) Phraseology and Second Language Proficiency. *Applied Linguistics*, **19**, 24-44. https://doi.org/10.1093/applin/19.1.24

[7] 潘璠. 语料库驱动的英语本族语和中国作者期刊论文词块结构和功能对比研究[J]. 外语与外语教学, 2016(4): 115-123.

[8] Conrad, S.M. and Biber, D. (2005) The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. *Lexicographica*, **20**, 56-71. https://doi.org/10.1515/9783484604674.56

[9] Biber, D., Conrad, S. and Cortes, V. (2004) If You Look at …: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, **25**, 371-405. https://doi.org/10.1093/applin/25.3.371

[10] Hyland, K. (2008) As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, **27**, 4-21. https://doi.org/10.1016/j.esp.2007.06.001

[11] 高霞. 基于中外学者学术论文可比语料库的词块使用研究[J]. 外语与外语教学, 2017(3): 42-52.

[12] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. and Hirst, G. (1999) The Longman Grammar of Spoken and Written English. Pearson Education Ltd., Harlow, 992.

[13] Cortes, V. (2004) Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes*, **23**, 397-423. https://doi.org/10.1016/j.esp.2003.12.001

[14] Chen, Y.-H. and Baker, P. (2010) Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology*, **14**, 30-49.

[15] Biber, D. and Barbieri, F. (2007) Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, **26**, 263-286. https://doi.org/10.1016/j.esp.2006.08.003

[16] Pan, F., Reppen, R. and Biber, D. (2016) Comparing Patterns of L1 versus L2 English Academic Professionals: Lexical Bundles in Telecommunications Research Journals. *Journal of English for Academic Purposes*, **21**, 60-71. https://doi.org/10.1016/j.jeap.2015.11.003

[17] 卫乃兴. 中国学习者英语口语语料库初始研究[J]. 现代外语, 2004, 27(2): 140-149.

[18] Hyland, K. (2012) Bundles in Academic Discourse. *Annual Review of Applied Linguistics*, **32**, 150-169. https://doi.org/10.1017/S0267190512000037